DOCUMENT RESUME

ED 418 131 TM 028 225

AUTHOR Tam, Hak P.; Li, Yuan H.

TITLE Is the Use of the Difference Likelihood Ratio Chi-square Statistic for Comparing Nested IRT Models Justifiable?

PUB DATE 1997-03-00

NOTE 27p.; Paper presented at the Annual Meeting of the American Educational Research Association (Chicago, IL, March 24-28, 1997).

PUB TYPE Reports - Evaluative (142) -- Speeches/Meeting Papers (150)

EDRS PRICE MF01/PC02 Plus Postage.

DESCRIPTORS *Chi Square; *Comparative Analysis; *Item Response Theory; Mathematical Models; Tables (Data)

IDENTIFIERS Difference (Concept); *Likelihood Ratio Tests; Nested Data

ABSTRACT

The main purposes of this study were to investigate, by means of simulation: (1) whether the difference likelihood ratio chi-square statistic (G2-dif) for comparing item response theory (IRT) models is asymptotically distributed as a chi-square distribution; and (2) the accuracy rate of applying G2-dif in selection of nested IRT models. Two hundred replications of simulated test data for 2 test lengths and 2 sample sizes were generated. The results of this study demonstrate that the usual practice of treating the G2-dif as distributed as a central chi-square distribution is not sound. For short test length, the proportion of times the correct model is being selected can be very low. It appears that the G2-dif is more likely to be distributed as a noncentral chi-square distribution. Discussion concerning the proportion of the right model is being selected by the difference statistic as well as its relative merits in comparison to the Akaike Information Criteria (AIC) and the m(k) indices is also included in this study. (Contains 12 tables and 26 references.) (Author/SLD)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

Is the Use of the Difference Likelihood Ratio Chi-square Statistic for Comparing Nested IRT Models Justifiable?

U.S. DEPARTMENT OF EDUCATION
Office of Educational Research and Improvement
EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)
This document has been reproduced as received from the person or organization originating it.
Minor changes have been made to improve reproduction quality.
Points of view or opinions stated in this document do not necessarily represent official OERI position or policy.

Hak P. Tam*
University of Maryland
National Taiwan Normal University

Yuan H. Li**
University of Maryland
Prince George's County Public Schools, Maryland

PERMISSION TO REPRODUCE AND DISSEMINATE THIS MATERIAL HAS BEEN GRANTED BY
Yuan Li
TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

ED 418 131

Paper presented at the annual meeting of the
American Educational Research Association
March 24-28, 1997, Chicago, IL.

* Now at the National Taiwan Normal University

** Graduate program at the University of Maryland and employed with the Prince George's County Public Schools, Maryland

TM028225

BEST COPY AVAILABLE

Is the Use of the Difference Likelihood Ratio Chi-square Statistic for Comparing Nested IRT Models Justifiable?

Abstract

The main purposes of this research are to investigate, by means of simulation, (a) whether the difference likelihood ratio chi-square statistic, G^2_{dif}, for comparing IRT models is asymptotically distributed as a chi-square distribution and (b) the accuracy rate of applying G^2_{dif} in selection of nested IRT models. The results based on this study demonstrate that the usual practice of treating the G^2_{dif} as distributed as a central chi-square distribution is not sound. For short test length, the proportion of times the correct model is being selected can be very low. It appears that the G^2_{dif} are more likely to be distributed as a noncentral chi-square distribution. Discussion concerning the proportion of the right model being selected by the difference statistic as well as its relative merits in comparison to the AIC and the m_k indices is also included in this study.

KEY WORDS: Difference Likelihood Ratio Chi-square Statistic, Difference Chi-square Statistic, Item Response Theory (IRT), BILOG, Likelihood Ratio Chi-square Statistic, Pearson Chi-square Statistic.

I. Introduction

The issue of model-data fit is of major concern when item response theory (IRT) models are used for practical testing. More specifically, the concern is whether or not test items fit the model assumed by practitioners (see, e.g., McKinley & Mills, 1985; Reise, 1990; Rogers, 1987; Smith, 1991; Yen, 1981). It is of course desirable to have most items fit the assumed model. If not, especially in the presence of numerous mis-fit items, the issue of model choice is critical. Currently, identifying the most parsimonious test model that retains the integrity of the observed data is an important motivation behind item-fit studies (Yen, 1981). However, the issue of selecting an appropriate IRT model has received less attention than the study of item fit. Poor model choice can lead to inaccurate conclusions of item dregging, as well as inappropriate assessment of differential item functioning.

To what extent can a certain IRT model be used to model a given set of examinees' responses to a test? A common method is to use the likelihood ratio chi-square goodness-of-fit statistic to measure the degree of data-model fit, as is also the case in latent class analysis and structural equation modeling, etc.. It is generally assumed that the likelihood ratio chi-square statistic is asymptotically distributed as a chi-square distribution with appropriate degrees of freedom as specified by the model. However, this statistic is not completely valid. This is because, in view of the numerous possible combination in response patterns, the frequency count of examinees in many response patterns will be very sparse, thereby violating the assumption of chi-square statistic that requires most of the expected frequencies be at least equal to five (see, e.g., Bock & Aitkin, 1981; Gitomer & Yamamoto, 1991; Reiser, VandenBerg, 1994).

Furthermore, a distinct characteristic in the IRT framework is that item parameter estimates derived from the joint maximum likelihood estimation may not be consistent as sample size and the number of items increase. This is because the abilities of the examinees are unknown and must be estimated along with item parameters (refer to Baker, 1992 for detailed discussion). Item parameters estimated by marginal maximum likelihood method (Bock & Aitkin, 1981) do not depend on the direct estimation of examinees' abilities, but rather on their ability distribution

(Baker, 1992). Thus, the estimate of the likelihood ratio chi-square statistic could be affected by special characteristics of the item parameter estimates.

Traditionally, the difference or component chi-square (G^2_{dif}) is used for the comparison of the relative fit of various IRT models with different parameter restrictions (e.g. Bock & Aitkin, 1981; Fischer & Parzer, 1991; Gitomer & Yamamoto, 1991), as well as for the assessment of differential item functioning (Camilli & Shepard, 1994; Thissen, Steinberg & Wainer,1993). This statistic is basically a ratio of the likelihood ratio chi-square statistic derived from the compact model to that derived from the general or subsuming model. Numerically, it is computed as the change in likelihood ratio chi-square statistics between a pair of hierarchically related models. Based on the additivity property of the likelihood ratio chi-square, G^2_{dif} is usually presumed to be asymptotically distributed as a chi-square distribution, with its degrees of freedom equal to the difference of the degrees of freedom between the two corresponding models. Yet as discussed above, the likelihood ratio chi-square statistic corresponding either to the subsuming or the nested model may not be chi-square distributed in the first place. Hence, the overall question raised in the present study is whether the use of the difference likelihood ratio chi-square statistic for comparing hierarchically nested IRT models (i.e. one model is a constrained form of the other) valid?

More specifically, the main purposes of this research are to investigate, by means of simulation, (a) whether the difference likelihood ratio chi-square statistic for comparing IRT models is asymptotically distributed as a chi-square distribution and (b) the accuracy rate of applying G^2_{dif} in selection of nested IRT models. Discussion concerning the proportion of the right model being selected by the difference statistic as well as its relative merits in comparison to the AIC (Bozdogan, 1987) and the m_k indices (McDonald & Mok, 1995) will also be included in this study.

Presented are a brief review of background theory in section two, a description of the methodology in section three, the results and discussion in section four, and the conclusion in section five.

BEST COPY AVAILABLE

II. The Likelihood Ratio Chi-square for Model Comparison

A. Likelihood Function of IRT Models

Under the three-PL logistic model (see Baker, 1992; Hambleton & Swaminathan, 1985; Mislevy & Bock, 1990), the probability, P_{ij}, of a correct response to the i^{th} item for the j^{th} examinee with ability θ_j is given by:

$$P_{ij}(\theta_j) = c_i + (1 - c_i)\frac{e^{Da_i(\theta_j - b_i)}}{1 + e^{Da_i(\theta_j - b_i)}} \qquad (1)$$

where a_i is the item discrimination, b_i is the item difficulty, c_i is the lower asymptote parameter (also known as the guessing parameter), and D (usually equal to 1.702) is a scaling factor. A two-PL model is attained if the guessing parameter c_i is constrained to zero for all items in (1) above. A one-PL model is a restricted form of the two-PL model by further constraining the item discrimination index a_i to be identical or equal to one for all items.

Assuming that the local independence assumption holds, given an examinee with ability θ who responds to a set of n items with the response pattern $\underline{u}$, then the probability of obtaining the response pattern $\underline{u}$ given θ and the item parameter vector $\underline{\xi}$ (a, b, c) can be computed by:

$$P(\underline{u}|\theta, \underline{\xi}) = \prod_{i=1}^{n} P_i^{u_i} Q_i^{1-u_i} \qquad (2)$$

where Q = 1 - P. If $\underline{\theta}$ is randomly sampled from a density distribution of ability $g(\underline{\theta})$, the unconditional probability is given by (see Baker, 1992; Mislevy & Bock, 1990):

$$P(\underline{u}, \theta|\underline{\xi}) = \prod_{i=1}^{n} P_i^{u_i} Q_i^{1-u_i} g(\theta) \qquad (3)$$

Then, the marginal probability of obtaining the response pattern $\underline{u}$ is obtained by integrating out the ability parameter $\underline{\theta}$ from the left side of (3), thereby giving:

$$P(\underline{u}|\underline{\xi}) = \int_{-\infty}^{\infty} P(\underline{u}|\theta, \underline{\xi}) g(\theta) d\theta = \pi_u \qquad (4)$$

BEST COPY AVAILABLE

As a result, the item parameters can be estimated without the estimation of the $\underline{\theta}$ parameters. This marginal probability of obtaining the response patterns $\underline{u}$, $P(\underline{u}|\underline{\xi})$, hereafter denoted by π_u, can be approximated to any desired degree of accuracy by the Gauss-Hermite quadrature via computing the sum:

$$\sum_{k}^{q} P(x = u \mid X_k) A(X_k) \qquad (5)$$

Here X_k represents a tabled quadrature point (node), and $A(X_k)$ is the corresponding weight which is related to the height of the density $g(\underline{\theta})$ in the neighborhood of the node X_k (see Stroud & Sechrest, 1966 for details).

Now let the subscript u represent a specific response pattern, r_u denote the number of examinees obtaining that specific response pattern, and s represent the number of distinct response patterns observed. In general, there are 2^n possible response patterns for n binary items, hence $s \leq 2^n$, ignoring those patterns with $r_u = 0$. Thus,

$$\sum_{u}^{S} r_u = N \qquad (6)$$

where N is the sample size. The likelihood function is then defined as the joint probability of all examinee's response patterns and is given by

$$L \propto \prod_{u=1}^{s} (\pi_u)^{r_u} \qquad (7)$$

Taking logarithm of Equation 7 results in

$$\ln L = k + r_u \sum_{u=1}^{s} \ln(\pi_u), \qquad (8)$$

where k is a constant which does not influence the estimation of the item parameters. The item parameter estimates are obtained by maximizing the log likelihood function presented in Equation 7. More specifically, they are obtained by differentiating ln L with respect to the item parameters a, b, and c, and solving the subsequent likelihood equations simultaneously. If the underlying shape of the ability distribution is correctly specified, the marginal maximum likelihood estimator can be consistent as the number of test items and the sample size increase (refer to Seong, 1990).

BEST COPY AVAILABLE

B. Testing the Goodness-of-fit of the IRT Model

When data from a large sample of examinees is available, the model-data fit may be tested either for the whole test or item by item. The likelihood ratio goodness-of-fit chi-square statistic for testing the assumed model against a general multinomial alternative is given by (Bock & Aitkin, 1981)

$$G^2 = 2\left(\sum_{u=1}^{s} r_u \ln \frac{r_u}{\pi_u N}\right) \quad . \qquad (9)$$

where G^2 is equal to -2 ln L as presented in Equation 8. This statistic may be asymptotically distributed as a chi-square distribution with s - mn -1 degrees of freedom (where m is the number of item parameters in the model).

However, if the number of all possible response patterns (2^n) is large relative to the sample size N, then most of the expected frequencies of the response patterns (π_u N) will be less than 5. This setting is quite common in practical testing situations. Bock and Aitkin(1981) suggested that the frequencies of response patterns with small expectations should be pooled until all expected frequencies equal or exceed 5. After pooling, the likelihood ratio chi-square statistic with s_p - mn -1 degrees of freedom (s_p is the number of response patterns after pooling), then provides a conservative test of data-model fit. Unfortunately, the likelihood function in Equation 8 has not actually been maximized in the pooled data (Bock & Aitkin, 1981). In addition, the way of pooling data is subjective and no IRT computer software at present can provide this kind of test when pooling data is necessary. Consequently, the fit of the model has seldom been assessed for a whole test in studies reported in the literature, except for the case of a short test with a large of number of examinees, such as the empirical dataset (5 test items by 1000 examinees) of the Law School Aptitude Test (LSAT) that were used in, among others, Bock and Lieberman (1970).

C. Model Selection in IRT Models

Regarding the choice of an appropriate IRT model, although the true model is never known, the usual practice is to determine if some models fit an observed dataset better than the others. One way to do this is to compute the likelihood ratio goodness-

BEST COPY AVAILABLE

of-fit chi-square statistics of the relevant models together with their associated degrees of freedom. But as explained above, the likelihood ratio chi-square statistics, in most cases, may not be appropriate for assessing data-model fit in IRT modeling. An often used alternative approach is to compare the relative fit of the two IRT models. Let G^2_g be the likelihood ratio chi-square statistic of a more general IRT model, while G^2_c is the corresponding statistic of a more constrained or nested version of the former model. The statistic used to assess the improvement in fit of the augmented model over the compact model is by the difference chi-square which can be expressed as:

$$G^2_{dif} = G^2_c - G^2_g = -2(\ln L_c) + 2(\ln L_g) \qquad (10)$$

where $\ln L_c$ and $\ln L_g$ can be derived from Equation 8. (Note that even for the same dataset, different IRT computer programs may handle either the constant k or the metric of the item parameters differently. Otherwise, they may employ different estimation algorithms. Thus the value of the -2 log likelihood reported in the computer output may differ from program to program).

It is important to notice that several assumptions have to be satisfied in order for the G^2_{dif} statistic to be approximately distributed as a chi-square distribution with its degrees of freedom equal to the degrees of freedom of the nested model minus that of the subsuming model. They are, among others, (a) the two models should be hierarchically related, (b) the fundamental IRT assumptions have been met in the estimation of both models, and (c) the more general of the two models provides a more proper specification for the data (see Holt & Macready, 1988).

Other properties of the difference chi-square statistic have been pointed out by Steiger, Shapiro & Browne (1985) in their seminal paper. Specifically, they demonstrated that the asymptotic intercorrelations among the chi-square statistics calculated for hierarchically related models on the same dataset can be quite high. However, the intercorrelations between the chi-square statistics and the sequential chi-square statistics computed from pairs of nested models should be asymptotically independent of each other. Also, the correlations among the difference chi-squares should be independent of each other.

BEST COPY AVAILABLE

Besides the difference chi-square, alternative model selection procedures are also pursued in this paper for comparison. A brief description is included here for handy reference. The first one is by means of Akaike's information criterion, or AIC, which is defined as (see Bozdogan, 1987)

$$AIC = -2 \ln L + 2m, \tag{11}$$

with m denoting the number of parameters estimated by an IRT model. The model of choice is usually regarded as the one that yields the lowest AIC value. This is by virtue of its definition; AIC penalizes the more complicated models in favor of the more parsimonious models. This index is included in the present study because its performance in terms of selecting an IRT model is not very well known. Another alternative is the m_k index which was originally suggested in the context of structural equation modeling (McDonald, 1988), but later introduced by McDonald and Mok (1995) as a measure of the goodness-of-fit for IRT models. It is defined as follows:

$$m_k = e^{-\frac{1}{2}(d_k)}, \tag{12}$$

where d_k is, in turn, defined as $(G^2 - df)/N$. Here d_k is actually a measure of the non-centrality parameter. The index m_k has the property that its values are scaled within the range from zero to one, and that the larger its value, the better the fit of the corresponding model. Since this is a relatively new statistic, its performance is not quite known and is thus included for investigation in the present study. Unfortunately, BILOG (Mislevy, Bock, 1990) does not report G^2 for tests longer than 10 items; hence the study of the performance of m_k is confined to the test length =5 situation only.

BEST COPY AVAILABLE

III. Methodology

In this section, an overview of the research design is first described followed by a discussion of the specific details as well as some explanation of the rationale.

A. Overview

Two hundred replications of simulated test data for two test lengths (items =5 and 50) in combination with two sample sizes (N=1000 and 2000) were generated from an existing item bank according to the 1-PL, 2-PL and 3-PL models. The combination of 5 items and 1000 subjects has been used in a number of studies and serves as a base for comparison. With knowledge of the true model underlying each data set, other IRT models were fitted to the data and the corresponding difference likelihood ratio chi-square statistic was calculated. These difference chi-square statistics were then assessed to see if they were approximately distributed as a chi-square distribution. Also, the intercorrelations of the various chi-square statistics were computed.

The likelihood ratio chi-square statistic is usually computed after item parameter estimates that will maximize the likelihood function are chosen. But as discussed earlier, the characteristics of the underlying $\underline{\theta}$ distribution may affect the estimation of the item parameters and may result in an incorrect estimation of the likelihood ratio chi-square statistic. In order to minimize this problem, the marginal maximum likelihood estimation procedure was used in this study to estimate the item parameters, while assuming that the underlying $\underline{\theta}$ distribution was assumed to be known for the 5-item test. The ability distribution for the 50-item test were empirically estimated. All analyses were performed using the BILOG software.

B. The Simulation of Test Data

Test Length (5, 50): The items together with their item parameters used to generate the two simulated tests in this study were selected from an existing Math Item Bank. At present it contains about 220 test items and was constructed by one of the public schools on the Eastern shore. First, 5 items were randomly selected to form the 5-item test. In practice, a 5-item test is too short to precisely measure an examinee's

BEST COPY AVAILABLE

ability. There were two reasons for constructing the 5-item test. The first reason was that the expected frequency for each possible response pattern would probably be larger than 5, thereby meeting the requirement of the chi-square test. In this case, the total possible number of response patterns would amount to 32 ($=2^5$). The expected frequency for each pattern would then be 31.25 for a sample size of 1000 examinees. Another reason was to follow the tradition of a long line of research, where two sections of the Law School Aptitude Test (LSAT), each with 5 items, was first studied by Bock & Lieberman (1970) and later re-analyzed by Bock & Aitkin (1980), Bartholomew (1980), Christoffersson (1975), McDonald & Mok (1995), and Muthen (1978). Thus, a 5-item test was included in the present study and to serve as a base for comparison of the appropriateness of applying the likelihood ratio chi-square test for model selection.

Next to produce the 50-item test, an additional 45 items were randomly selected from the item bank and combined with the five previously selected items. Tables 1 and 2 present some descriptive statistics of the item parameters used to simulate the short and the longer tests. Notice that when test-score data were simulated according to the one-PL model later on, the values of 1.0 and 0.0 were assigned for all the discrimination indices a_i and the guessing parameters c_i, respectively. For the two-PL datasets, the values of 0.0 were used for all the guessing parameters.

Table III-1
Descriptive Statistics of the Item Parameters for the 5-Item Test

	a		b		c	
Model	Mean	Range	Mean	Range	Mean	Range
One-PL	1.00	1.00 to 1.00	-0.70	-1.66 to 0.29	0.00	0.00 to 0.00
Two-PL	0.80	0.60 to 1.03	-0.70	-1.66 to 0.29	0.00	0.00 to 0.00
Three-PL	0.80	0.60 to 1.03	-0.70	-1.66 to 0.29	0.13	0.06 to 0.20

BEST COPY AVAILABLE

Table III-2
Descriptive Statistics of the Item Parameters for the 50-Item Test

	a		b		c	
Model	Mean	Range	Mean	Range	Mean	Range
One-PL	1.00	1.00 to 1.00	-0.40	-2.50 to 3.00	0.00	0.00 to 0.00
Two-PL	0.90	0.30 to 1.40	-0.40	-2.50 to 3.00	0.00	0.00 to 0.00
Three-PL	0.90	0.30 to 1.40	-0.40	-2.50 to 3.00	0.16	0.04 to 0.31

Ability and Sample Sizes (1000, 2000): In this study, the ability parameters were randomly selected from the standard normal distribution, N(0,1). First, 1000 ability parameters were selected and used for N=1000 datasets. Then, for N=2000 datasets, an additional 1000 ability parameters were selected and combined with the previous ability parameters of the 1000-sample size datasets. This way of constructing ability parameters has previously been used by McKinley and Mills (1985). These ability parameters were held constant across the 200 replications of data under each combination of study conditions. The reason for this decision was to retain the same metric for the estimated item parameters across the 200 replications for the test length=5 situation. (But see the discussion in the Calibration and Analysis subsection below for the test length=50 situation). Furthermore, since the likelihood ratio chi-square was calculated after the estimation of the item parameters, the above procedure retained the same metric for the likelihood ratio chi-square statistic across the 200 replications within any specific study condition.

Simulation of Datasets: The probability of each examinee answering an item correctly was computed according to Equation (1) or the like depending on the underlying IRT model. Uniform random numbers in the interval [0,1] were then generated and compared with the examinees' probabilities of success. If the probability was larger than the corresponding generated random number, the examinee was scored 1, otherwise the examinee was scored 0. A total of twelve combinations of conditions (two test lengths X two sample sizes X three IRT models) were considered in this study. Two hundred replications were generated under each condition.

BEST COPY AVAILABLE

C. Calibration and Analysis

All likelihood ratio chi-square statistics were computed by using BILOG. Two options in BILOG deserve further explanation.

The first one is the FREE option, which when adopted, will instruct the program to empirically estimate the θ distribution of the respondents. Otherwise, the default is to assume the ability parameter to be distributed as a unit normal. In the present study, this option was invoked for the test length=50, but not for the test length=5 situation. This was because for a short test, the empirical posterior of the distribution of the ability parameter may not be accurate. Hence, the ability distribution was assigned the default unit normal distribution, which is the same as the underlying distribution of the simulated dataset. For a longer test, the empirical posterior can be quite accurately estimated, and so the FREE option was adopted. Moreover, when FREE was used for the test length=50, owing to sampling fluctuation, the estimated posterior ability distribution might be a little bit different from replication to replication. Consequently, the metric of the item parameters might also be different from replication to replication. Strictly speaking, item linking should be performed. However, such differences should be minimal as each dataset was generated from the identical set of ability parameters which were held constant across replications across each combination of study condition. Finally, in a real life situation, the ability distribution is actually unknown. The FREE option was used to estimate the latent ability distribution.

The second one is the FLOAT option. If this option is adopted, the means of the item parameter prior distributions will be estimated along with the item parameters (see Mislevy & Bock, 1990). Otherwise, the means of the item parameters distribution will be fixed at their default values during the estimation process. In this study, FLOAT was invoked for test length=50, but not for test length=5.

With the knowledge of the true model behind each dataset, three IRT models (1-PL, 2-PL, and 3-PL) were fitted to each of them and the corresponding likelihood ratio chi-square statistic assessed. Two hundred likelihood ratio chi-square statistics were then separately obtained for the replications within each combination of designed conditions. Afterward, the following analyses were conducted:

(1). In order to determine if the difference chi-square statistic was really chi-square distributed, the distributions of the observed G^2_{dif} statistics were examined by comparing them to the theoretical central chi-square distributions with the appropriate degrees of freedom. Following Holt & Macready (1989), each distribution of 200 observed G^2_{dif} statistics was classified into 10 intervals as defined by the set of 0th, 10th, ... 90th, and the 100th quantiles of the central chi-square distribution with 9 degrees of freedom. Each of the intervals will, therefore, contain an expected frequency count of 20. A Pearson chi-square statistic with 9 degrees of freedom was computed to assess the fit of the observed G^2_{dif} to a central chi-square distribution.

(2). In addition to the overall fit, each of the observed G^2_{dif} distributions was examined by comparing its observed mean and standard deviation with its corresponding expected mean and standard deviation.

(3). The intercorrelations among the likelihood ratio chi-square statistics computed for each model as well as their relationship with the various difference chi-squares computed from pairs of hierarchically related models were examined.

(4). The proportion of times when the true models were correctly chosen over the attempted models by the difference chi-square, the AIC and the m_k indices were also computed for comparison purposes.

BEST COPY AVAILABLE

IV. Results and Discussion

A. The Distribution of the Difference Likelihood Chi-square Test

The observed distributions of the G^2_{dif} for each combination of study condition were first examined by comparing the observed mean and standard deviation with their expected values. If the difference statistics were really distributed as a chi-square distribution, then for test length=50, the theoretical mean of the difference chi-square between the 1-PL and the 2-PL models would be 50 and the corresponding standard deviation would be 10 (S.D. = $\sqrt{2df} = \sqrt{100}$). The values for the difference between the 1-PL and 3-PL, and between the 2-PL and 3-PL were similarly calculated. The results are listed in Table IV-1. As seen there, the observed means and standard deviations were not close to the corresponding expected values. It was especially the case for the situation when sample size=2000, and when simpler models were fitted to datasets that were generated by more complicated models.

Table IV-1
Comparisons of the observed means and standard deviations of difference chi-square statistics, G^2_{dif}, with their expected values when test length=50 (Sample Sizes = 1000, 2000; Replications = 200)

Model Comparison		True Model					
		One-PL		Two-PL		Three-PL	
	N	1000	2000	1000	2000	1000	2000
1 vs 2	M (50.00)	45.97	30.60	655.36	1260.59	597.28	1159.30
	SD(10.00)	18.67	25.86	52.52	105.15	45.25	65.73
1 vs 3	M (100.0)	90.34	107.18	699.82	1342.49	706.14	1352.38
	SD(14.14)	19.81	34.57	50.99	103.95	50.71	72.93
2 vs 3	M (50.00)	44.38	76.58	44.46	81.90	108.86	193.08
	SD(14.14)	20.47	37.23	18.14	26.93	21.12	27.95

* Expected statistics are given in parentheses.

The distributions of the difference chi-square were then examined according to their overall fit to central chi-square distributions with 9 degrees of freedom. The results are presented in Table IV-2 below. All the goodness-of-fit tests were statistically significant at the .05 level, and hence none of the distributions of the observed difference chi-square were distributed as a central chi-square distribution.

BEST COPY AVAILABLE

Similar to Table IV-1, Table IV-2 indicates that the goodness-of-fit was worse under a larger sample size, and when the true model was more complicated than the attempted model.

Table IV-2
The Pearson chi-square statistics for assessing the fit of the observed G^2_{diff} statistics to a central chi-square distribution when test length=50 (Sample Sizes =1000, 2000; Replications = 200)

	True Model					
	One-PL		Two-PL		Three-PL	
	Sample Size N					
	1000	2000	1000	2000	1000	2000
Model Comparison						
1 vs 2	113.50	604.30	1800.00	1800.00	1800.00	1800.00
1 vs 3	148.20	179.10	1800.00	1800.00	1800.00	1800.00
2 vs 3	291.30	746.50	162.80	888.30	1663.70	1800.00

For test length=5, the observed and the expected means and standard deviations of the difference chi-square between pairs of nested models are presented in Table IV-3. The observed descriptive statistics were not close to the corresponding expected values. When the 3-PL was attempted to fit to datasets generated either from the 1-PL or the 2-PL model, the means of the difference chi-square statistics took on a negative value, which is rather unusual.

Table IV-3
Comparisons of the observed means and standard deviations of difference chi-square statistics, G^2_{diff}, with their expected values when test length=5 (Sample Sizes = 1000, 2000; Replications = 200)

Model Comparison		True Model					
		One-PL		Two-PL		Three-PL	
	N	1000	2000	1000	2000	1000	2000
1 vs 2	M (5.00)	4.25	3.78	13.39	23.66	7.57	10.21
	SD(3.30)	3.42	2.90	7.29	9.45	4.65	6.36
1 vs 3	M (10.0)	-1.71	-2.25	9.94	19.14	10.86	15.19
	SD(4.47)	6.07	5.83	7.97	10.30	6.25	8.97
2 vs 3	M (5.00)	-5.96	-6.02	-3.45	-4.52	2.29	4.99
	SD(3.30)	4.77	5.34	4.41	5.06	4.05	6.26

* Expected statistics are given in parentheses.

BEST COPY AVAILABLE

The distributions of the difference chi-square were then examined in exactly the same way as under test length=50. The goodness-of-fit statistics are presented in Table IV-4. The goodness-of-fit tests were again statistically significant at the .05 level, and the distributions of the observed difference chi-square under test length=5 were not distributed as central chi-square distributions either.

Table IV-4
The Pearson chi-square statistics for assessing the fit of the observed G^2_{dif} statistics to a central chi-square distribution when test length=5 (Sample Sizes =1000, 2000; Replications = 200)

	True Model					
	One-PL		Two-PL		Three-PL	
	Sample Size N					
	1000	2000	1000	2000	1000	2000
Model Comparison						
1 vs. 2	45.30	46.60	765.80	1682.30	141.10	309.60
1 vs. 3	336.30	1387.80	95.30	555.90	44.50	260.60
2 vs. 3	1531.20	1495.00	1301.60	1370.30	137.60	152.80

One possible reason why the difference statistics were not distributed as central chi-square distributions may parallel the study by Holt and Macready (1989) in the context of latent class analysis. In both situations, the more parsimonious models (e.g. 1-PL, 2-PL) were obtained from the subsuming model (the 3-PL in this study) by constraining some parameters (here the guessing parameters) to their boundary values (zero in this study). Hence, a regularity condition was violated and the difference statistics may not have been chi-square distributed.

As regards the anomaly of obtaining negative difference chi-squares under marginal maximum likelihood estimation for test length=5, with prior distribution of θ fixed (see Table IV-3), the problem may be related to the fact that this test is basically an easy test (see Table III-1). But in this study, the θ distribution was fixed at N(0,1), so there were relatively fewer low ability parameter values generated. Hence, the guessing parameters were not appropriately estimated as there were few observations available, thus rendering their standard errors very large. Under these

BEST COPY AVAILABLE

circumstances. Thissen & Wainer (1982) indicated that "the large covariance between lower asymptote and location (difficulty) then causes this uncertainty to move partially to the estimate of location. With more difficult items the effect is lessened somewhat. The two-parameter model has problems as well, but they are far less severe" (Thissen & Wainer, 1982, pp. 403-404). For this reason, both the difficulty and the guessing parameters were not accurately estimated under the 3-PL. The final model may not "fit" better than the 1-PL and 2-PL models, thereby producing negative difference chi-square statistics. This was especially the case when the datasets were generated from the 1-PL model. The problem of negative difference statistics did not occur in test length=50 because the item parameters were quite accurately estimated. Also for a longer test, the estimate of the marginal probability of a response pattern is more reliable and accurate.

B. The Proportions of Times the True Models Were Correctly Chosen

The proportions of times the true models were chosen over the attempted models by the difference chi-square and AIC for test length=50 under various situations is presented in Table IV-5. Selection of nested models by the m_k index is not considered for the reason explained in section 2 above.

Although the difference chi-square statistics were not central chi-square distributed, their performance in terms of the proportions of times the true models were correctly chosen over the attempted model were relatively high, ranging from .885 to 1.00 for sample size=1000. However, when the sample size=2000 and when the 3-PL was attempted to fit data generated either from the 1-PL or the 2-PL, the proportions of times the true models were correctly chosen can be quite low, ranging from .33 to .725. As seen from Table IV-5 regarding the performance of AIC, the proportions of times the true models were correctly chosen were quite high, especially when the true models were less complicated than the attempted models. This is basically consistent with the knowledge that AIC penalizes the more complicated model.

BEST COPY AVAILABLE

Table IV-5
The proportion of times the true models were chosen over the attempted models by the difference chi-square and AIC for test length = 50 (Sample Sizes=1000, 2000; Replications = 200)

Model Comparison	True Model					
	One-PL		Two-PL		Three-PL	
N	1000	2000	1000	2000	1000	2000
1 vs 2	0.885	0.915	1.000	1.000		
1 vs 3	0.955	0.725			1.000	1.000
2 vs 3			0.900	0.325	0.965	1.000
AIC	0.995	0.990	0.990	0.725	0.675	1.000

The results for test length=5 are presented in Table IV-6. Here the difference chi-squares were less likely to identify the true models, especially when the true models were more complicated than the attempted models. The performance of AIC was even worse when the underlying models were more complicated models. Based on Table IV-6, the m_k index appeared to prefer models in this order: 2-PL, 1-PL, then 3-PL, regardless of sample sizes.

Table IV-6
The proportion of times the true models were chosen over the attempted models by the difference chi-square, AIC and m_k for test length = 5 (Sample Sizes =1000, 2000; Replications = 200)

Model Comparison	True Model					
	One-PL		Two-PL		Three-PL	
N	1000	2000	1000	2000	1000	2000
1 vs. 2	0.945	0.970	0.585	0.845		
1 vs. 3	1.000	1.000			0.125	0.320
2 vs. 3			0.990	1.000	0.030	0.135
AIC	1.000	1.000	0.000	0.000	0.000	0.000
m_k	0.685	0.725	0.870	0.940	0.285	0.425

Basic descriptive statistics for the m_k index are presented in Table IV-7 below. Obviously, most of the values were very close to the upper limit, and the spread of values was extremely small.

BEST COPY AVAILABLE

Table IV-7
The descriptive statistics of Mean and Standard Deviation for m_k index (Test Length = 5; Sample Sizes=1000, 2000; Replications = 200)

Attempted Model	True Model: One-PL				Two-PL				Three-PL			
	N: 1000		2000		1000		2000		1000		2000	
	M	SD	M	SD	M	SD	M	SD	M	SD	M	SD
One-PL	.999	(.004)	1.000	(.002)	.994	(.005)	.995	(.003)	.997	(.004)	.998	(.002)
Two-PL	.998	(.003)	.999	(.002)	.999	(.004)	1.000	(.002)	.998	(.004)	.999	(.002)
Three-PL	.993	(.003)	.997	(.002)	.994	(.004)	.997	(.002)	.997	(.003)	.999	(.001)

In addition, for test length=5, the likelihood ratio goodness-of-fit chi-square statistic for testing the attempted model against a general multinomial alternative can be computed according to Equation 9. The proportions of times the attempted models were identified to fit the datasets by the likelihood ratio chi-square are presented in Table IV-8. When the true model was the 1-PL, it can be computed from the table that the Type I error rates of the goodness-of-fit test in identifying the true model amounted to 0.12 and 0.09, for sample size=1000 and 2000 respectively. Likewise, when the true model was the 2-PL, the corresponding Type I error rates were 0.11 and 0.10 for the two respective sample sizes. Lastly, when the true model was the 3-PL, the corresponding Type I error rates jumped to 0.265 and 0.175 respectively. In all cases, the probability of committing Type I error by the goodness-of-fit test statistics exceeded the 0.05 nominal alpha rate. In addition, the values off the diagonal in Table IV-8 denote the probabilities of committing Type II error under various situations. As can be seen there, the power of the likelihood ratio chi-square test can be quite low.

Table IV-8
The proportions of times the attempted models were identified to fit the datasets by the likelihood ratio chi-square goodness-of-fit test (Test Length = 5; Sample Sizes =1000, 2000; Replications = 200)

Attempted Model	True Model: One-PL		Two-PL		Three-PL	
N	1000	2000	1000	2000	1000	2000
One-PL	0.880	0.910	0.620	0.290	0.750	0.645
Two-PL	0.865	0.855	0.890	0.900	0.845	0.845
Three-PL	0.300	0.340	0.455	0.480	0.735	0.825

BEST COPY AVAILABLE

C. The Intercorrelation Matrix among Chi-square Statistics.

The intercorrelations among the various likelihood ratio chi-squares as well as with the difference chi-squares computed from pairs of nested models when the test length=50 are provided in Table IV-9. The numbers in parentheses are correlations under sample size=2000, while those without parentheses are correlations under sample size=1000 situation.

Table IV-9
The intercorrelations among the various likelihood ratio chi-squares as well as with the difference chi-squares computed from pairs of nested models when the test length=50 (Sample Sizes =1000, 2000; Replications =200)

True Model	Attempted Model	One	Two	Three	D12	D23	D13
_1	One	1.0(1.0)	.99(.99)	.99(.99)	.01(-.01)	.14(.02)	.16(.02)
_2		1.0(1.0)	.97(.94)	.97(.94)	.19(.19)	-.08(.05)	.17(.21)
_3		1.0(1.0)	.97(.97)	.96(.96)	.16(.08)	.04(.06)	.16(.09)
_1	Two		1.0(1.0)	.99(.99)	-.08(-.09)	.18(.06)	.12(-.00)
_2			1.0(1.0)	.99(.99)	-.05(-.16)	-.02(.11)	-.06(-.13)
_3			1.0(1.0)	.99(.99)	-.08(-.16)	.04(.04)	-.05(-.13)
_1	Three			1.0(1.0)	-.03(-.04)	.08(-.06)	.06(-.09)
_2				1.0(1.0)	-.03(-.14)	-.10(.02)	-.07(-.14)
_3				1.0(1.0)	-.08(-.17)	-.08(-.06)	-.10(-.18)
-1	D12				1.0(1.0)	-.49(-.45)	.44(.27)
-2					1.0(1.0)	-.26(-.17)	.94(.97)
-3					1.0(1.0)	.04(.06)	.91(.92)
-1	D23					1.0(1.0)	.57(.74)
-2						1.0(1.0)	.09(.08)
-3						1.0(1.0)	.45(.44)
-1	D13						1.0(1.0)
-2							1.0(1.0)
-3							1.0(1.0)

In the table, datasets that were actually generated by the 1-PL, 2-PL and 3-PL models were given the numerical labels _1, _2, and _3, respectively. The verbal labels ONE, TWO, and THREE represent the attempted models to fit the data were the 1-PL, 2-PL and the 3-PL, respectively. For example, the number .99 in the first row and second column of the matrix represents a strong positive linear correlation between the likelihood chi-square statistics produced by fitting the 1-PL to datasets generated by the 1-PL model with those produced by fitting the 2-PL model to the same datasets. Likewise, the number .97 in the second row and second column represents a strong positive linear correlation between the likelihood chi-squares

BEST COPY AVAILABLE

produced by fitting the 1-PL to datasets generated by the 2-PL model with those produced by fitting the 2-PL model to the same datasets.

The label D12 represents the difference chi-square statistics produced by fitting the 1-PL and 2-PL to datasets generated by the same model. Consider, for example, the number .01 in the first row and fourth column of the matrix. Here the underlying true model is the 1-PL, so .01 denotes no linear correlation between the likelihood chi-squares produced by fitting the 1-PL with the difference chi-squares derived from fitting the 1-PL and 2-PL to the same datasets.

Apparently, the upper left quadrant of the matrix indicates that, regardless of what the true model was, the intercorrelations among the likelihood chi-squares derived from various models were very high. The correlations between the likelihood chi-squares and the various difference chi-squares in the upper right quadrant were quite weak, indicating that they were independent from each other. One reason behind this observation is that, using 1-PL and 3-PL for illustration, the G^2 values for the 3-PL models are larger than those computed for the 1-PL models for some datasets, while smaller for the other datasets.

Finally, the intercorrelations among the differences chi-squares could be quite high, which is different to those found in Steiger et al. (1985). It should be pointed out, however, that the theorems in Steiger et al. were stated in relation to noncentral chi-square distributions.

The results for test length=5 are presented in Table IV-10 below. Basically, the same patterns were found conformable to the previous table, except perhaps the upper right quadrant. There the correlations were moderately high in some instances.

Table IV-10

The intercorrelations among the various likelihood ratio chi-squares as well as with the difference chi-squares computed from pairs of nested models when the test length=5 (Sample Sizes =1000, 2000; Replications =200)

True Model		Attempted Model One	Two	Three	D12	D23	D13
_1	One	1.0(1.0)	.90(.93)	.65(.73)	.55(.33)	.31(.16)	.55(.31)
_2		1.0(1.0)	.70(.61)	.63(.50)	.74(.82)	.02(.13)	.69(.81)
_3		1.0(1.0)	.84(.76)	.69(.42)	.54(.73)	.36(.42)	.63(.81)
_1	Two		1.0(1.0)	.73(.77)	.12(-.04)	.33(.21)	.32(.17)
_2			1.0(1.0)	.81(.73)	.05(.04)	.18(.35)	.14(.21)
_3			1.0(1.0)	.83(.50)	-.01(.11)	.41(.61)	.26(.49)
_1	Three			1.0(1.0)	.06(.02)	-.40(-.47)	-.28(-.41)
_2				1.0(1.0)	.13(.10)	-.42(-.38)	-.12(-.09)
_3				1.0(1.0)	-.03(.11)	-.17(-.39)	-.13(-.20)
-1	D12				1.0(1.0)	.07(-.10)	.62(.41)
-2					1.0(1.0)	-.14(-.09)	.84(.87)
-3					1.0(1.0)	.02(.01)	.76(.72)
-1	D23					1.0(1.0)	.83(.87)
-2						1.0(1.0)	.43(.41)
-3						1.0(1.0)	.67(.71)
-1	D13						1.0(1.0)
-2							1.0(1.0)
-3							1.0(1.0)

BEST COPY AVAILABLE

V. Conclusion

All in all, based on the above examination and discussion, it is clear that the usual practice of treating the difference chi-square as distributed as a central chi-square distribution is not sound. For short test length, the proportion of times the correct model is being selected can be very low. It appears that the difference chi-squares are more likely to be distributed as a noncentral chi-square distribution. Hence a natural extension of the present study is to estimate the noncentral parameter and then test if the difference statistic is distributed as a noncentral chi-square with appropriate degrees of freedom.

So far as the performance of the selection indices in the context of IRT is concerned, both the AIC and the m_k indices are not very satisfactory. A promising index, namely, root mean square error of approximation (RMSEA), has recently drawn the attention of researchers in structural equation modeling (Steiger, 1980; McDonald & Mok, 1995). This index has not been pursued in the present study due to the fact that some of the G^2 values were less than their corresponding degrees of freedom. Apparently, more work needs to be done in the area of model selection within an IRT context.

BEST COPY AVAILABLE

References

Baker, F. B. (1992). Item response theory: Parameter estimation techniques. New York: Marcel Dekker, Inc.

Bartholomew, D. J. (1980). Factor analysis for categorical data. Journal of the Royal Statistical Society, 42, 293-321.

Bock, R. D. & Aitkin, M. (1981). Marginal maximum likelihood estimation of item parameters: application of an EM algorithm. Psychometrika, 46, 443-459.

Bock, R. D. & Lieberman, M. (1970). Fitting a response model for n dichotomously scored items. Psychometrika, 35, 179-196.

Bozdogan, H. (1987). Model Selection and Akaike's information criterion (AIC): The general theory and its analytical extensions. Psychometrika, 3, 345-370.

Camilli, G. & Shepard, L. A. (1994). Methods for identifying biased test items. Thousands Oaks, CA: SAGE Publications, INC.

Christoffersson, A. (1975). Factor analysis of dichotomized variables. Psychometrika, 40, 5-32.

Fischer, G. H. & Parzer, P. (1991). An extension of the rating scale model with an application to the measurement of change. Psychometrika, 56, 637-651.

Gitomer, D. & Yamamoto, K. (1991). Performance modeling that integrates latent trait and class theory. Journal of Educational Measurement, 28, 173-189.

Hambleton, R. K. & Swaminathan, H. (1985). Item response theory: Principles and applications. Boston: Kluwer Nijhoff Publishing.

Holt, J. A. & Macready, G. B. (1989). A simulation study of the difference Chi-square statistic for comparing latent class models under violation of regularity conditions. Applied Psychological Measurement, 13, 221-231.

Kinnucan, M. T. & Wolfram, D. (1990). Direct comparison of bibliometric models. Information Processing & Management, 26, 777-790.

McDonald, D. P. (1988). An index of goodness-of-fit based on noncentrality. Journal of Classification, 6, 97-103.

McDonald, R. P. & Mok, Magdalena M. -C. (1995). Goodness of fit in item response models. Multivariate Behavioral Research, 30, 23-40.

BEST COPY AVAILABLE

McKinley, R. L. & Mills C. N. (1985). A comparison of several goodness-of-fit statistics. Applied Psychological Measurement, 9, 49-57.

Mislevy, R. J. & Bock, R. D. (1990). BILOG 3: Item analysis and test scoring with binary logistic models. Mooresville, IN: Scientific Software, Inc.

Muthen, B. (1978). Contribution to factor analysis of dichotomous variables. Psychometrika, 43, 551-560.

Reise, S. P. (1990). A comparison of item- and person-fit methods of assessing model-data fit in IRT. Applied Psychological Measurement, 14, 127-137.

Reiser, M. & VandenBerg, M. (1994). Validity of the chi-square test in dichotomous variable factor analysis when expected frequencies are small. British Journal of Mathematical and Statistical Psychology, 47, 85-107.

Rogers, H. J. (1987). A Monte Carlo investigation of several person and item fit statistics for item response models. Applied Psychological Measurement, 11, 47-57.

Seong, T. (1990). Sensitivity of marginal maximum likelihood estimation of item and ability parameters to the characteristics of the prior ability distributions. Applied Psychological Measurement, 14, 299-311.

Smith, R. M. (1991). The distributional properties of Rasch item fit statistics. Educational and Psychological Measurement, 51, 541-565.

Steiger, J. H., Shapiro, A. & Browne, M. W. (1985). On the multivariate asymptotic distribution of sequential chi-square statistics. Psychometrika, 50, 253-264.

Stroud, A. H., & Sechrest, D. (1966). Gaussian quadrature formulas. Englewood Cliffs (N.J.): Prentice-Hall.

Thissen, D., Steinberg, L. & Wainer, H. (1993). Detection of differential item functioning using the parameters of item response models. In P. W. Holland and H. Wainer (Ed.), Differential item functioning (pp. 67-113). Hillsdale, NJ: Educational Testing Service.

Yen W. M. (1981). Using simulation results to choose a latent trait model. Applied Psychological Measurement, 5, 245-262.